Mostly Searchable: Piles of Paper → Digital Files

What I had

Years of paid bills and receipts

sitting in boxes

getting in the way

impossible to find anything

What I wanted

Searchable archive

not too much effort

not too much outlay

some of my storage space back

Hardware

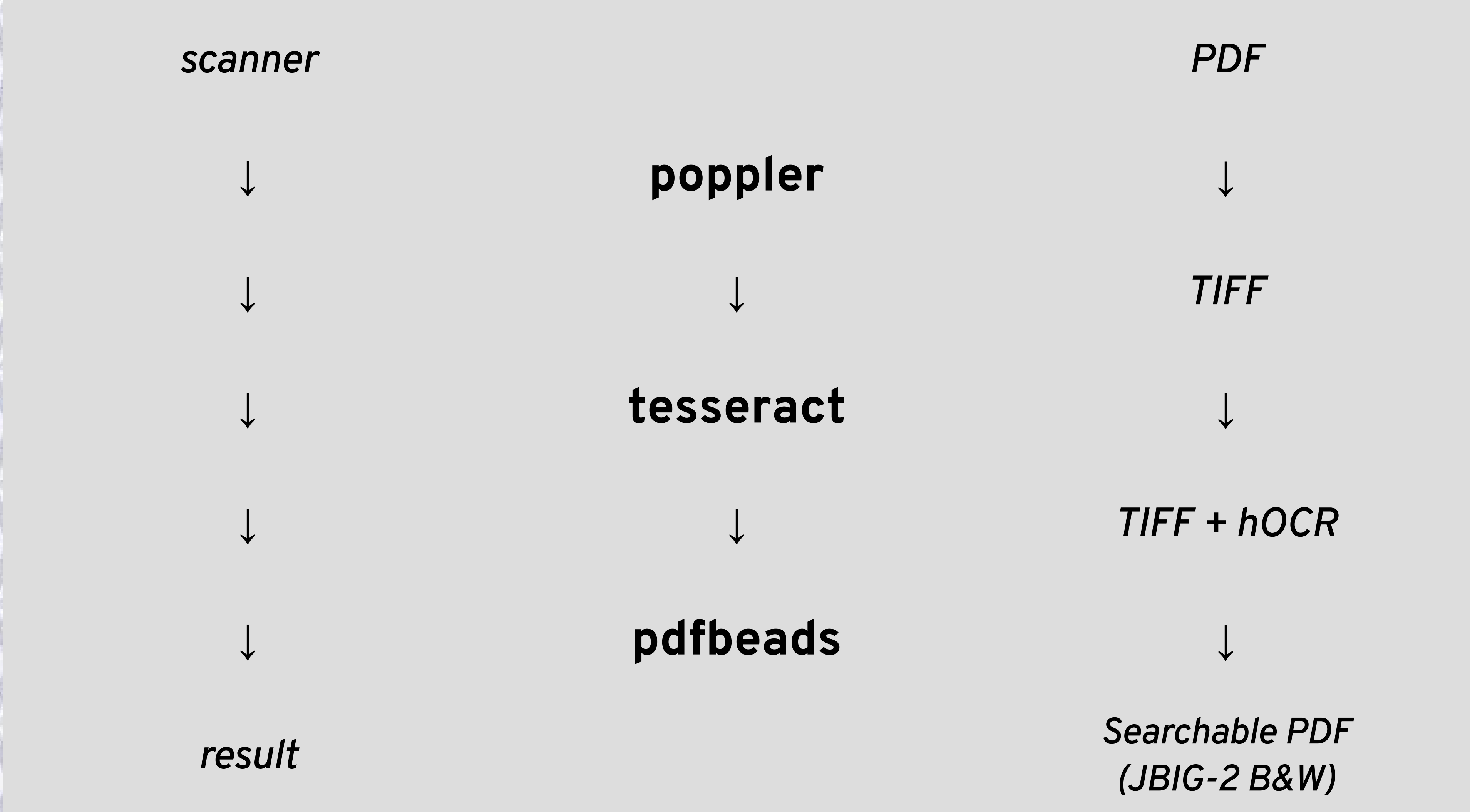
Epson WF-7520

duplex scanner

scan to SMB-shared SD card

size of a small 🚍

Software



scanner		*PDF*
↓	**poppler**	↓
↓	↓	*TIFF*
↓	**tesseract**	↓
↓	↓	*TIFF + hOCR*
↓	**pdfbeads**	↓
result		*Searchable PDF* *(JBIG-2 B&W)*

Searching / Indexing

Unity Dash Search
recoll
tracker
beagle

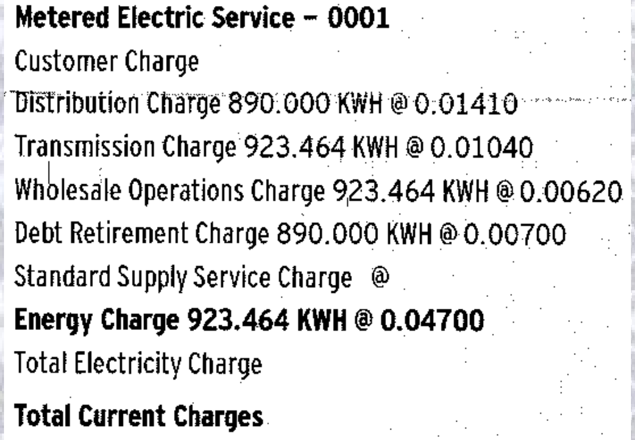
...

Mac OS Spotlight
Windows Search

...

(yeah, I cheated here)

Results



Metered Electric Service - 0001

Customer Charge

Distribution Charge 890.000 KWH @ 0.01410

Transmission Charge 923.464 KWH @ 0.01040

Wholesale Operations Charge 923.464 KWH @ 0.00620

Debt Retirement Charge 890.000 KWH @ 0.00700

Standard Supply Service Charge @

Energy Charge 923.464 KWH @ 0.04700

Total Electricity Charge

Total Current Charges

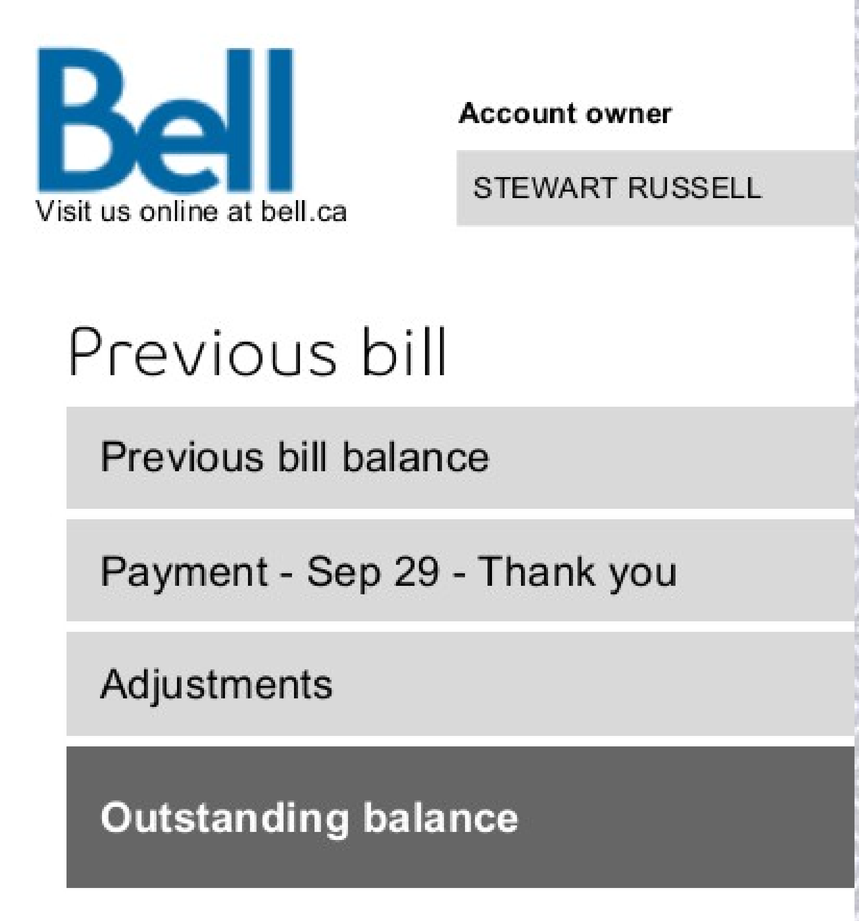
```
Metered Electric Service -0001
Customer Charge
Translnission Charge 923.464 KWHCa0.01040
Wholesale Operations Charge 923.464
Debt Retirement Charge 890.000
KWHCm0.00620
KWHCa0.00700
Standard Supply Service Charge
Energy Charge 923.464 KWNLa0.04700
Total Electricity Charge
Total Current Charges
```

10 ~ 50 KB/page

page image readable

search mostly okay

A Whole New Horror



Bell

Visit us online at bell.ca

Account owner

STEWART RUSSELL

Previous bill

Previous bill balance

Payment - Sep 29 - Thank you

Adjustments

Outstanding balance

Online bills

huge PDFs

~ 250 K/page

embedded logos
embedded fonts

… makes me sad and tired.

Possible Solution

Rasterize to 200 dpi greyscale

OCR surprisingly accurate

convert via ⚠JPEG-2000⚠ to PDF

30 ~ 50 KB / page

work in progress …